Thinking About Thinking and Its Cognitive Limits[*]

Adam Brandenburger[†] Xiaomin Li[‡]

Version 08/30/15

Abstract

Evidence shows that when engaged in game-playing tasks, individuals think about what other individuals are thinking. Presumably, though, cognitive limits prevent individuals from entertaining indefinite levels of thinking about what other players are thinking about what other players are thinking, etc.. Drawing on neural evidence, we propose a procedure by which a player builds such levels of thinking. The number of possibilities that a player must consider at each level m in this procedure grows exponentially with m. We argue that this feature may help explain why a cognitive bound has been found in game-theory studies to come into effect at a small finite number of levels of thinking about thinking.

These particular uncertainties — as to the other player's beliefs about oneself — are almost universal, and it would constrict the application of a game theory fatally to rule them out.
– Daniel Ellsberg [10, 1959]

1 Introduction

When one first hears about game theory, a basic question comes quickly to mind: How does a player Ann in a game think not just about what moves another player Bob might make, but also about what Bob might be thinking about her own moves, and, perhaps, about still higher levels of thinking about thinking? Yet the historical development of game theory sidestepped this issue. In **minimax theory** (von Neumann [32, 1928], von Neumann and Morgenstern [33, 1944]), players adopt a worst-case rather than predictive view of what other players do, and choose accordingly. In **equilibrium theory** (Nash [25, 1951]), each player is assumed to have access to the actual strategies chosen by the other players and to choose a strategy accordingly. Players do not have to operate on the basis of guesses about other players. A more recent development, **epistemic game theory**, is different in making "thinking about what another player is thinking" basic to the analysis of games; we will come back to this approach later.

Outside game theory, iterated thinking has been extensively studied in **cognitive psychology** and **cognitive neuroscience**. Here, the term used is **Theory of Mind** (**ToM**), defined

[*] Our thanks to Samson Abramsky, Jessy Hsieh, Natalya Vinokurova, and Jiemai Wu for valuable comments, to Elliot Lipnowksi for a very important suggestion, and to the NYU Stern School of Business and NYU Shanghai for financial support.

[†] Stern School of Business, Polytechnic School of Engineering, NYU Shanghai, New York University, New York, NY 10012, U.S.A., adam.brandenburger@stern.nyu.edu, adambrandenburger.com

[‡] Behavioral & Social Neuroscience Program, California Institute of Technology, Pasadena, CA 91125, U.S.A., xli2@caltech.edu

as the ability to think about another person's beliefs, wants, and intentions (Singer and Tusche [28, 2014, p.514]), and it is well accepted that people possess and use ToM in many situations.

Of particular significance for game theory is evidence from behavioral and brain-imaging studies showing that regions of the brain active in ToM processing are also activated when people play a game with a human (but not with a computer) counterpart (McCabe et al. [22, 2001], Gallagher and Frith [14, 2003]). Other important evidence comes from Sally and Hill [27, 2006], who examined how children, some developing normally and some with autistic spectrum disorder (ASD), played a number of games. (In pioneering work, Baron-Cohen, Leslie, and Frith [3, 1985] showed that children with ASD perform considerably less well on ToM tasks.) In the repeated Prisoner's Dilemma, Sally and Hill found that better performance on ToM tasks was correlated with higher levels of cooperation. In the Ultimatum Game, ASD proposers offered the responder less than other proposers did, as compared with the Dictator Game, where the two groups made the same offers. To the extent that behavior involving more cooperation (Prisoner's Dilemma) or more generosity (Ultimatum Game) is more effective for a player, these findings connect ToM ability to strategic ability.

There appear to be good empirical — not just intuitive — grounds for making the examination of "thinking about what another player is thinking" an important part of game theory.

2 Theory-of-Mind Ability

A person's **ToM ability** is usually defined via a task where the person hears a short story describing a social situation and is then asked questions about the story (Kinderman, Dunbar, and Bentall [19, 1998], Stiller and Dunbar [30, 2007]). The questions differ in terms of the number of levels of "Ann thinks Bob thinks Charlie thinks . . . " that they contain (Ann, Bob, and Charlie are characters in the story). The maximum number of such levels that a question can contain and still be answered correctly by the person gives that person's ToM ability.

There is considerable empirical evidence that people's ToM abilities are limited to a small finite number of levels. Employing a **narrative approach**, Stiller and Dunbar [30, 2007] found that the modal level at which subjects failed ToM questions was level 5. (The subjects were normal adults.[1]) Two recent papers in **experimental game theory** found iterated thinking up to level 3 and level 4, respectively. Arad and Rubinstein [1, 2012] introduced a game (they call it the "11-20 game") which they designed to prompt iterated thinking and to permit robust identification of levels. They found that a model allowing up to level-3 thinking (together with some noise) fitted their experimental data best. Kneeland [20, 2015] introduced a novel experimental design (using "ring games") that allows one to identify levels of thinking in a whole family of games, while making much weaker assumptions on behavior and beliefs than in the previous experimental games literature. In her experiments, Kneeland found that 6 percent of subjects were level-0 (i.e., made dominated choices), 23 percent were level-1, 27 percent were level-2, 22 percent were level-3, and 22 percent were level-4.

The findings in Kneeland [20, 2015], in particular, point to a somewhat higher degree of cognitive capability concerning thinking about thinking than was typically found in the earlier games literature; see Nagel [24, 1995], Stahl and Wilson [29, 1995], Costa-Gomes, Crawford, and Broseta [8, 2001], Camerer, Ho, and Chong [7, 2004], and others. This earlier — and pioneering — literature employed more tightly specified models (the "level-k" and "cognitive hierarchy" models) in order to achieve identification, which could mean that higher-level thinkers are under-identified.

[1] Women achieved one more level (an average of 5.53 levels) than did men (an average of 4.41 levels), a difference which was significant.

The broad agreement on number of levels across Stiller and Dunbar [30, 2007], Arad and Rubinstein [1, 2012], and Kneeland [20, 2015] is encouraging. We do note, though, that some care is needed in comparing across narrative-based and game-based studies. In the first type of study, a statement of the exact form "Ann thinks Bob wants to play Left" is not employed, but similar statements are used, and they are coded as level 2. In game-based studies, if Ann is observed to choose a strategy that is best for her when Bob chooses Left, then she would be coded as at least a level-1 player (she optimizes relative to some belief about what Bob does), but not necessarily higher.[2] So, there is a question as to whether the Stiller and Dunbar [30, 2007] limit of level 4 'translates' to a limit of level 4 or level 3 in a game context. Another distinction is that Stiller and Dunbar [30, 2007] find a range of limits across subjects, with a significant number of subjects succeeding in tasks requiring 5 or more levels of thinking. This may reflect the greater degree of priming involved in asking subjects to read narratives which point rather directly to what one character thinks another character is thinking, as compared with asking subjects to play games.

In sum, we believe there is good evidence of a **cognitive limit** on thinking about thinking that comes into operation at a small finite number of levels. A limit of three to four levels receives some good empirical support.[3] No doubt, this is a soft not hard upper bound, influenced by many factors including contextual ones as just mentioned, presence or absence of memory aids, auxiliary incentives, and more. Nevertheless, we will use this limit as a broad guideline in the following sections.

3 Models of Thinking about Thinking

Harsanyi [15, 1967-8] was first in game theory to recognize that uncertainty on the part of players about some aspect of the **structure** of a game (a common example is uncertainty about other players' payoff functions) would naturally lead players to form hierarchies of beliefs. In such a hierarchy, a particular player has a first-order belief (over the uncertainty in question), a second-order belief (over other players' first-order beliefs), and so on to higher orders. More recently, epistemic game theory has developed to address uncertainty on the part of players about the **strategies** chosen in a game; see Brandenburger [5, 2014, Introduction] for an overview. Here, too, players are assumed to form hierarchies of beliefs — this time, over strategies in the game rather than over structure of the game.

These advances improve the descriptive power of game theory in that they enable us, as game theorists, to include **epistemic** components of a strategic situation that traditionally were left out of a game description. With this richer description, game theory can now ask and answer questions of the form: "If the players in a game think such-and-such (including what they think others think, and so on), then what play(s) of the game will be observed?" But these advances do not seem, in themselves, to help in understanding a process by which players might build hierarchies of beliefs. It is useful to look at why this is so.

Suppose there is a space of underlying uncertainty S, which we take to be finite, as depicted in Figure 1. (This space might describe uncertainty about structure or strategies or both.) Define the space of first-order beliefs on S to be the space $\mathcal{M}(S)$ of probability measures on S. (This

[2] In particular, if Left is a bad choice for Bob — a dominated choice, say — then Ann would be coded as precisely a level-1 player.

[3] Keynes, always the intuitive thinker, alighted on a similar number of levels in his famous analogy between the stock market and a beauty contest [18, 1936]: "It is not a case of choosing those [faces] which, to the best of one's judgment, are really the prettiest, nor even those that average opinion genuinely thinks the prettiest. We have reached the third degree where we devote our intelligences to anticipating what average opinion expects the average opinion to be. And there are some, I believe, who practice the fourth, fifth and higher degrees."

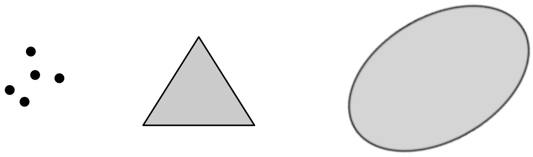

Figure 1: The sets S, $\mathcal{M}(S)$, and $\mathcal{M}(\mathcal{M}(S))$ (resp.)

is often identified with the $(|S|-1)$-dimensional simplex.) Note that $\mathcal{M}(S)$ is an infinite space, so to define second-order beliefs on S, i.e., to define probability measures on $\mathcal{M}(S)$, we need to put a measure-theoretic structure on $\mathcal{M}(S)$. In probability theory, a standard notion of a well-behaved infinite space X is that it is Polish.[4] This works well for defining hierarchies of beliefs, because the space $\mathcal{M}(X)$ is again Polish.[5] In our case, the space $\mathcal{M}(\mathcal{M}(S))$ will then be well-defined (Polish), and the same will be true of all higher-order spaces of beliefs.

This construction gives us a way to describe hierarchies of beliefs, but it does not seem very satisfactory as a belief formation process. Suppose we envisage a player as scanning each belief space (first-order, second-order, etc.) and choosing a belief to hold from each. If we count cases at each level, then this process would involve high complexity even at the level of first-order beliefs, but then no further increase in complexity. This is because the space $\mathcal{M}(S)$ of possible first-order beliefs is already uncountably infinite, but then all higher-order spaces are no larger. (A standard fact: All uncountably infinite Polish spaces have the cardinality $2^{\aleph_0}$ of the continuum.[6]) This jump and then lack of further increase in complexity does not fit well with the empirical evidence reviewed in Section 2.

Admittedly, case counting is a very crude measure of complexity of a space of probability measures. Very likely better would be some measure based on the complexity of search in an infinite space. But, rather than stipulate such a measure, we will instead review some neural evidence on belief formation in games reported in Bhatt and Camerer [4, 2005] and build a — very simple — model of belief formation from this evidence.

4 Neural Evidence on Belief Formation

Bhatt and Camerer [4, 2005] conducted an fMRI study of subjects engaged in playing a number of matrix games. (The games are taken from an earlier study by Costa-Gomes, Crawford, and Broseta [8, 2001].) Players were asked to make choices and to state first- and second-order beliefs about strategy choices in the games.[7] Point beliefs were elicited, and we will use the word "thinks" as a shorthand for such beliefs. Write $\hat{s}_a$ (resp. $\hat{r}_a$) for the proposition that Ann chooses strategy s_a (resp. r_a), write $\hat{s}_b$ (resp. $\hat{r}_b$) for the proposition that Bob chooses strategy s_b (resp. r_b), and write $\Box_a$ (resp. $\Box_b$) for the modal operator "Ann thinks" (resp. "Bob thinks").[8]

[4] A Polish space is a topological space which is separable and completely metrizable (Kechris [17, 1995, p.13]).

[5] Formally, $\mathcal{M}(X)$ is the space of all probability measures on the Borel σ-algebra of X, endowed with the weak topology. See Kechris op.cit., p.68, pp.109-110, and Theorem 17.23 for definitions and the proof that $\mathcal{M}(X)$ is Polish. This method of constructing hierarchies of beliefs comes from Brandenburger and Dekel [6, 1993].

[6] Kechris op. cit., Corollary 6.5.

[7] The games were presented to the players in a 'transparent' way, so it is assumed that players were not uncertain about the structure of the game.

[8] We will make no formal use of modal logic in this paper, but it will be convenient to adopt some modal-logic notation.

Three events concerning a player (here, Ann) were studied in Bhatt and Camerer [4, 2005]:

$$\hat{s}_a \tag{1}$$

$$\square_a \, \hat{s}_b \tag{2}$$

$$\square_a \, \square_b \, \hat{r}_a \tag{3}$$

Brain activity of a player in event (3) was found to have more similarities with activity in event (1) than with activity in event (2). That is, there was more similarity at the neural level between choosing a strategy and forming a second-order belief than between forming a first-order belief and forming a second-order belief. Also, significantly greater activation was found in the **anterior insula** region of a player's brain in event (3) than in event (2). The anterior insula is part of the insular cortex, itself part of the cerebral cortex. It is associated with, inter alia, subjective feelings, attention, and, particularly relevant here, cognitive choices and intentions (Craig [9, 2009]).[9]

A priori, one might have expected that the neural processes involved in forming first-order and second-order beliefs would be most similar, and that choice behavior would be most distinct neurally. The above findings are surprising in this respect, and they prompt Bhatt and Camerer [4, 2005] to formulate what they call the self-referential strategic thinking hypothesis: Second-order belief formation is a combination of belief-formation and choice-making processes.

5 An Anchoring and Adjusting Process

The Bhatt and Camerer [4, 2005] hypothesis suggests one way in which we could put structure on a belief formation process. Rather than inspecting spaces of beliefs as a whole and picking specific beliefs from them, Ann might select a candidate strategy choice and then examine her view as to whether or not Bob thinks she intends to make this choice. This sets up the dichotomy: (i) Ann intends to play strategy s_a and thinks Bob thinks this; or (ii) Ann intends to play strategy s_a and it is not the case that she thinks Bob thinks this. Ann repeats this process for the various different strategy choices she considers from her overall set of possible strategies.

We can call an intended strategy choice for Ann an **anchor**, with the help of which she is able to think about what another player might be thinking about her choice. She can stay with this anchor, or **adjust** away from it, according as cases (i) or (ii) above obtain. There is precedent for proposing anchoring and adjusting processes in the ToM literature, albeit not in the particular context or form used here. See, for example, Epley et al. [11, 2004] and Tamir and Mitchell [31, 2010]. The latter paper describes an fMRI study where individuals were asked to report their own preferences and also their judgments about another individual's preferences. It specifically tested the hypothesis that judgments further away from the anchor entail greater cognitive processing than do closer judgments, and found supporting evidence. (It would be interesting to look for an analogous result in the context of game playing, with an appropriate metric on strategy sets.)

To specify this process further, we have to be more precise about case (ii) above. We also need to consider higher-order beliefs. We will undertake these steps in the next section. Beforehand, let us summarize the steps in this section. The main step is to suppose that a player's candidate strategy choices are used to reduce a large space of beliefs to two parts — one part associated with agreement with a candidate choice and another associated with disagreement. At the same time, we took the step of restricting the **modality of belief** for a player to be a point belief.

[9] The anterior insula has, in addition, been hypothesized to play a fundamental role in human awareness; see Craig op.cit..

(Remember that "thinks" is our term for the point-belief modality.) An objection at this point could be that the move to point beliefs is enough in itself to address the complexity explosion we saw in Section 3. The use of an anchoring and adjusting process seems superfluous. But our use of point beliefs is meant only as a first and very crude approximation to a more satisfactory approach that employs a belief modality in-between the fully probabilistic and one based on point beliefs. Moreover, we have seen that the anchoring and adjusting process has some empirical basis, so we want to build a model — albeit a very preliminary one — with this ingredient.

6 Epistemic Equilibrium and Disequilibrium

As Bhatt and Camerer [4, 2005] observe, agreement vs. disagreement between Ann's choice and what she thinks Bob thinks she chooses can be thought as an **equilibrium** vs. **disequilbrium** distinction. The **epistemic** view of Nash equilibrium from Aumann and Brandenburger [2, 1995] turns out to be well-suited to elaborating this point. We first review [2, 1995].

Say Ann is **rational** if $\hat{s}_a$ and $\Box_a \hat{r}_b$ imply that s_a maximizes Ann's (expected) payoff when she assigns probability 1 to Bob's choosing r_b. Define rationality for Bob similarly (with Ann and Bob interchanged). Here is a first set of epistemic conditions for Nash equilibrium, based on [2, 1995]:[10]

$$\hat{s}_a \tag{4}$$
$$\hat{s}_b \tag{5}$$
$$\Box_a \hat{s}_b \tag{6}$$
$$\Box_b \hat{s}_a \tag{7}$$
$$\text{Ann is rational} \tag{8}$$
$$\text{Bob is rational} \tag{9}$$

These conditions immediately imply that the strategy pair (s_a, s_b) must constitute a Nash equilibrium. Here is a second set of epistemic conditions, again based on [2, 1995]:

$$\Box_a \hat{s}_b \tag{10}$$
$$\Box_b \hat{s}_a \tag{11}$$
$$\Box_a \Box_b \hat{s}_a \tag{12}$$
$$\Box_b \Box_a \hat{s}_b \tag{13}$$
$$\Box_a [\text{Bob is rational}] \tag{14}$$
$$\Box_b [\text{Ann is rational}] \tag{15}$$

It is easy to see that the strategy pair (s_a, s_b) must again constitute a Nash equilibrium.[11]

The conditions in [2, 1995] were stated for Nash equilibrium as usually defined, i.e., as an **inter-player** concept. But they can be immediately modified to yield an **intra-player** concept, which is the application we need. To do this, we 'subjectivize' the conditions, with respect

[10] We remind the reader that we are making very 'soft' use of modal logic. In particular, conditions (8) and (9) could be more formally stated but we will not need to do so.

[11] It is a triviality that these sets of epistemic conditions yield Nash equilibrium. Indeed, conditions (4)-(9) were termed merely a Preliminary Observation in [2, 1995] for this reason. The purpose of the observation was to dispel a widespread impression in the literature that play of a Nash equilibrium somehow requires infinite-order ("common") knowledge or belief on the part of the players. Conditions (10)-(15) are based on Theorem A in [2, 1995], which gives epistemic conditions for mixed strategies to constitute a Nash equilibrium. The conditions trivialize here because we treat only point beliefs.

to Ann, say. By this we mean that we put the "Ann thinks" operator $\Box_a$ in front of each of the conditions (4)-(9) (and, subsequently, the conditions (10)-(15)). We adopt the axioms: $\hat{s}_a \rightarrow \Box_a \hat{s}_a$, $\Box_a \hat{s}_b \rightarrow \Box_a \Box_a \hat{s}_b$, and [Ann is rational] $\rightarrow \Box_a$ [Ann is rational]. (These axioms say that Ann can introspect about her own intentions and beliefs.) We then get:

$$\hat{s}_a \tag{16}$$
$$\Box_a \hat{s}_b \tag{17}$$
$$\Box_a \Box_b \hat{s}_a \tag{18}$$
$$\text{Ann is rational} \tag{19}$$
$$\Box_a \text{[Bob is rational]} \tag{20}$$

These conditions are now subjective, because they refer only to Ann's own action and epistemic state. They say that Ann thinks that she and Bob play the Nash equilibrium (s_a, s_b). The conditions constitute an **internal epistemic equilibrium**.

Internal epistemic **disequilibrium** arises when conditions (16)-(20) do not jointly hold, which raises the question of which of the conditions should be changed. At this point, we follow the route suggested by the neural evidence in Bhatt and Camerer [4, 2005] and change Ann's second-order belief (condition (18)) relative to her choice (condition (16)). Since the beliefs are point beliefs, we next have to decide how we will implement the idea of 'taking the negation of a point belief.' We do not see clear empirical guidance here, and we restrict ourselves to a kind of symbolic exploration around the base condition (18).

We first add negation symbols to (18) in the seven possible distinct ways:

$$\neg \Box_a \Box_b \hat{s}_a \tag{21}$$
$$\Box_a \neg \Box_b \hat{s}_a \tag{22}$$
$$\Box_a \Box_b \neg \hat{s}_a \tag{23}$$
$$\neg \Box_a \neg \Box_b \hat{s}_a \tag{24}$$
$$\Box_a \neg \Box_b \neg \hat{s}_a \tag{25}$$
$$\neg \Box_a \Box_b \neg \hat{s}_a \tag{26}$$
$$\neg \Box_a \neg \Box_b \neg \hat{s}_a \tag{27}$$

These conditions are easier to read when we push the negation symbol in. Write $\Diamond_a$ (resp. $\Diamond_b$) for the modal operator "Ann (resp. Bob) considers it possible that". Then the rewrite rule $\neg \Box_a$ implies $\Diamond_a \neg$ (likewise for Bob) yields:

$$\Diamond_a \Diamond_b \neg \hat{s}_a \tag{28}$$
$$\Box_a \Diamond_b \neg \hat{s}_a \tag{29}$$
$$\Box_a \Box_b \neg \hat{s}_a \tag{30}$$
$$\Diamond_a \Box_b \hat{s}_a \tag{31}$$
$$\Box_a \Diamond_b \hat{s}_a \tag{32}$$
$$\Diamond_a \Diamond_b \hat{s}_a \tag{33}$$
$$\Diamond_a \Box_b \neg \hat{s}_a \tag{34}$$

In words, these conditions say, respectively: (28) Ann considers it possible that Bob considers it possible that Ann does not choose s_a; (29) Ann thinks that Bob considers it possible that Ann does not choose s_a; (30) Ann thinks that Bob thinks that Ann does not choose s_a; (31) Ann considers it possible that Bob thinks that Ann chooses s_a; (32) Ann thinks that Bob considers it possible that Ann chooses s_a; (33) Ann considers it possible that Bob considers it possible that Ann chooses s_a; and (34) Ann considers it possible that Bob thinks that Ann does not choose s_a.

Adopting the usual axiom that $\Box_a$ implies $\Diamond_a$ (if Ann thinks a proposition is true then she certainly considers it possible) and the analogous axiom for Bob, we see that conditions (31), (32), and (33) are consistent with equilibrium. (They are implied by condition (18).) So, we use the term "disequilibrium" when any of other conditions — namely, (28), (29), (30), or (34) — hold. This is what we will mean by the negation of the equilibrium condition (18).

We summarize our belief formation process thus far. In forming a second-order belief, Ann considers five cases — one case of internal epistemic equilibrium and four cases of internal epistemic disequilibrium.

Continuing, we next subjectivize conditions (10)-(15), while adopting analogous introspection axioms to earlier. This yields:

$$\Box_a \, \hat{s}_b \tag{35}$$

$$\Box_a \, \Box_b \, \hat{s}_a \tag{36}$$

$$\Box_a \, \Box_b \, \Box_a \, \hat{s}_b \tag{37}$$

$$\Box_a \, [\text{Bob is rational}] \tag{38}$$

$$\Box_a \, \Box_b \, [\text{Ann is rational}] \tag{39}$$

We also assume an analogous belief formation process to earlier, now lifted from comparing (16) with (18), to comparing (35) with (37). There are fifteen possible ways to add negation symbols to (37), which, after rewriting and rejecting those propositions which are consistent with (37), reduce to the following eight cases:

$$\Diamond_a \, \Diamond_b \, \Diamond_a \, \neg \, \hat{s}_b \tag{40}$$

$$\Box_a \, \Diamond_b \, \Diamond_a \, \neg \, \hat{s}_b \tag{41}$$

$$\Box_a \, \Box_b \, \Diamond_a \, \neg \, \hat{s}_b \tag{42}$$

$$\Box_a \, \Box_b \, \Box_a \, \neg \, \hat{s}_b \tag{43}$$

$$\Diamond_a \, \Box_b \, \Box_a \, \neg \, \hat{s}_b \tag{44}$$

$$\Diamond_a \, \Box_b \, \Diamond_a \, \neg \, \hat{s}_b \tag{45}$$

$$\Box_a \, \Diamond_b \, \Box_a \, \neg \, \hat{s}_b \tag{46}$$

$$\Diamond_a \, \Diamond_b \, \Box_b \, \neg \, \hat{s}_b \tag{47}$$

In forming a third-order belief, Ann considers nine cases: one case of internal epistemic equilibrium and eight cases of internal epistemic disequilibrium.

7 Complexity of Thinking About Thinking

We can now summarize our model of belief formation. Ann tentatively fixes a strategy pair (s_a, s_b). Relative to this, she considers, in forming an mth-order belief, for $m \geq 2$, a total of $2^m + 1$ cases — one case of internal epistemic equilibrium, and 2^m cases of internal epistemic disequilibrium. Specifically, for her second-order belief, Ann considers five cases, for her third-order belief, she considers nine cases, for her fourth-order belief, she considers seventeen cases, and so on. This is the process we envisage for each candidate strategy pair (s_a, s_b) Ann considers, and it is repeated for different candidates from the overall set of possible strategy pairs.

Given the number of assumptions and simplifications we have made, the specific progression we find from five cases to nine cases to seventeen cases (from second-order to third-order to fourth-order beliefs) should not be taken literally. Nevertheless, we think there are grounds for cautious optimism in relating our numbers to empirical findings. Qualitatively, we get an **exponential increase** in the number of cases a player has to consider at each level. This finding of an exponential — rather than linear or polynomial — dependence of number of cases on m fits well with the empirical finding that a cognitive limit on levels of thinking sets in at small m. To go further and look for quantitative agreement between a model like ours and cognitive limits on levels of thinking, we suspect that it will be necessary to include additional ingredients in the analysis. Consideration of **working memory** capacity seems very likely to be important. We do not expect this to be a simple next step, given the shift away from a view that there is a fixed number of items that can be stored in working memory to a view that there is a tradeoff between number of items and precision of recall (Ma, Husain, and Bays [21, 2014]).

8 Discussion

An important extension to the game experiments on which we based our assumptions will be to collect information on players' higher-order beliefs (beyond the second-order beliefs reported in Bhatt and Camerer [4, 2005]). In Section 6, we extrapolated the neural evidence relating choice and second-order beliefs, to hypothesize an analogous relationship between first-order beliefs and third-order beliefs (conditions (35) and (37)). Clearly, direct evidence is needed to support such an extrapolation.

A second extension to Bhatt and Camerer [4, 2005] will be to distinguish more clearly internal epistemic equilibrium (Section 6) from reasoning about **levels of rationality**. Take a two-by-two matrix where Ann has a dominant strategy s_a (Bob does not), and once Ann's dominated strategy is eliminated, Bob has a dominant strategy s_b in the remaining submatrix. Then the conditions (19), (20), and "$\Box_a \Box_b$ [Ann is rational]" together imply conditions (16), (17), and (18). Thus, a hypothesis on reasoning about levels of rationality implies the internal epistemic equilibrium hypothesis, which undercuts the justification for the second hypothesis. Clearly, in games with more than one (pure) Nash equilibrium, this problem does not arise, and it will be important to study a wide class of games including a variety of games of this kind.

Games with more than two players raise a number of immediate questions. The extension of the definition of internal epistemic equilibrium is straightforward, but one can envisage several ways in which the definition of internal epistemic disequilibrium is extended. For example, if Ann thinks that Bob does not think she chooses the strategy she intends to choose, will Ann necessarily think the same of Charlie? Or, can she, at the same time, think that Charlie does think she chooses the strategy she intends to choose? The number of different kinds of disequilibrium that are allowed will make a big difference to the number of cases that arise at each level of thinking. Neural monitoring in experiments with **n-player games** is needed to guide extensions to our two-player game model in Section 6.

Another question about model building concerns the best belief modality to use. We explained in Section 3 why we did not choose the probabilistic modality. Instead, we went to the 'opposite extreme' and chose a point belief modality. As already noted, we see this choice as just a starting point, and it will be important to study intermediate modalities.

In this paper, we depart significantly from more conventional game-theoretic analysis, relative both to equilibrium theory and to epistemic game theory. First, a comparison with **equilibrium theory**. Nash equilibrium has, of course, traditionally been understood as an inter-player concept, which brings in an element of objective correctness. This is immediately clear in conditions

(4)-(9) in Section 6, where Ann thinks that Bob chooses the strategy s_b that he actually does choose (and likewise with Ann and Bob interchanged). This objective correctness may be a good assumption in some contexts (e.g., as the outcome of a learning process), but, unsurprisingly, it has been found empirically invalid in many experiments. (See, among others, Flood [12, 1958] for early experimental evidence against Nash equilibrium, McKelvey and Palfrey [23, 1992] for a prominent test in the context of the Centipede Game (Rosenthal [26, 1982]), and Healy [16, 2011] for a direct test of the epistemic conditions for Nash equilibrium in Aumann and Brandenburger [2, 1995].)

Once Nash equilibrium is subjectivized, as in conditions (16)-(20), it refers only to an individual's own (subjective) state of mind and there is no longer any requirement of objective correctness. Moreover, we have seen that there is some evidence (Bhatt and Camerer [4, 2005]) that this notion of internal Nash equilibrium has empirical significance in understanding belief formation. The use of Nash equilibrium as an intra- rather than inter-player concept may be fruitful in other ways, too.

With respect to **epistemic game theory**, we have changed the usual line of analysis. The usual approach is to start with a model of hierarchies of beliefs of the kind reviewed in Section 3 and then, within this model, impose epistemic conditions of interest and deduce what behavior by the players is consistent with these conditions. This has worked as a deductive exercise, but seems less suited to building a model of belief formation. For this purpose, we built a model where a player entertains a candidate strategy choice and considers beliefs that lie in a certain relationship to this choice.

We emphasize that our model should be seen as highly tentative and, best case, as directionally correct in terms of what accounts for a cognitive limit on thinking about thinking. Also, our model applies, at least directly, to thinking about thinking in games and not in narratives or other settings. In all cases, more detailed model building is clearly needed, but, in this paper, we did not want to get ahead of the neural evidence as we understand it.

References

[1] Arad, A., and A. Rubinstein, "The 11-20 Money Request Game: A Level-k Reasoning Study," *American Economic Review*, 102, 2012, 3561-3573.

[2] Aumann, R., and A. Brandenburger, "Epistemic Conditions for Nash Equilibrium," *Econometrica*, 63, 1995, 1161-1180.

[3] Baron-Cohen, S., A. Leslie, and U. Frith, "Does an Autistic Child Have a 'Theory of Mind'?" *Cognition*, 21, 1985, 37-46.

[4] Bhatt, M., and C. Camerer, "Self-Referential Thinking and Equilibrium as States of Mind in Games: fMRI Evidence," *Games and Economic Behavior*, 52, 2005, 424-459.

[5] Brandenburger, A., *The Language of Game Theory: Putting Epistemics into the Mathematics of Games*, World-Scientific, 2014.

[6] Brandenburger, A., and E. Dekel, "Hierarchies of Beliefs and Common Knowledge," *Journal of Economic Theory*, 59, 1993, 189-198.

[7] Camerer, C., T. Ho, and J. Chong, "A Cognitive Hierarchy Model of Games," *The Quarterly Journal of Economics*, 119, 2004, 861-898.

[8] Costa-Gomes, M., V. Crawford, and B. Broseta, "Cognition and Behavior in Normal-Form Games: An Experimental Study," *Econometrica*, 69, 2001, 1193-1235.

[9] Craig, A., "How Do You Feel — Now? The Anterior Insula and Human Awareness," *Nature Reviews: Neuroscience*, 10, 2009, 59-70.

[10] Ellsberg, D., "Rejoinder," *Review of Economics and Statistics*, 16, 1959, 42-43.

[11] Epley, N., B. Keysar, L. Van Boven, and T. Gilovich, "Perspective Taking as Egocentric Anchoring and Adjustment," *Journal of Personality and Social Psychology*, 87, 2004, 327-339.

[12] Flood, M.,"Some Experimental Games," *Management Science*, 5, 1958, 5-26.

[13] Frith, C., "Schizophrenia and Theory of Mind," *Psychological Medicine*, 34, 2004, 385-389.

[14] Gallagher, H., and C. Frith, "Functioning Imaging of 'Theory of Mind'," *TRENDS in Cognitive Science*, 7, 2003, 77-83.

[15] Harsanyi, J., "Games with Incomplete Information Played by 'Bayesian' Players, I-III," *Management Science*, 14, 1967-8, 159-182, 320-334, 486-502.

[16] Healy, P., "Epistemic Conditions for the Failure of Nash Equilibrium," 2011, available at http://healy.econ.ohio-state.edu/papers/Healy-EpistemicConditions.pdf.

[17] Kechris, A., *Classical Descriptive Set Theory*, Springer-Verlag, 1995.

[18] Keynes, J., *The General Theory of Employment, Interest and Money*, Macmillan, 1936.

[19] Kinderman, P., R. Dunbar, and R. Bentall, "Theory-of-Mind Deficits and Causal Attributions," *British Journal of Psychology*, 89, 1998,191-204.

[20] Kneeland, T., "Identifying Higher-Order Rationality," 2015, forthcoming in *Econometrica*, available at http://terri.microeconomics.ca.

[21] Ma, W., M. Husain, and P. Bays, "Changing Concepts of Working Memory," *Nature Neuroscience*, 17, 2014, 347-356.

[22] McCabe, K., D. Houser, L. Ryan, V. Smith and T. Trouard, "A Functional Imaging Study of Cooperation in Two-Person Reciprocal Exchange," *Proceedings of the National Academy of Sciences*, 98, 2001, 11832-11835.

[23] McKelvey, R., and T. Palfrey, "An Experimental Study of the Centipede Game," *Econometrica*, 60, 1992, 803-836.

[24] Nagel, R., "Unraveling in Guessing Games: An Experimental Study," *American Economic Review*, 85, 1995, 1313-1326.

[25] Nash, J., "Non-cooperative Games," *Annals of Mathematics*, 54, 1951, 286-295.

[26] Rosenthal, R., "Games of Perfect Information, Predatory Pricing, and the Chain Store Paradox," *Journal of Economic Theory*, 25, 1982, 92-100.

[27] Sally, D., and E. Hill, "The Development of Interpersonal Strategy: Autism, Theory-of-Mind, Cooperation and Fairness," *Journal of Economic Psychology*, 27, 2006, 73-97.

[28] Singer, T., and A. Tusche, "Understanding Others: Brain Mechanisms of Theory of Mind and Empathy," in Glimcher, P., and E. Fehr, *Neuroeconomics: Decision Making and the Brain*, 2nd edition, Academic Press, 2014.

[29] Stahl, D. and P. Wilson, "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, 10, 1995, 218-254.

[30] Stiller, J., and R. Dunbar, "Perspective-Taking and Memory Capacity Predict Social Network Size," *Social Networks*, 29, 2007, 93-104.

[31] Tamir, D., and J. Mitchell, "Neural Correlates of Anchoring-and-Adjustment During Mentalizing," *Proceedings of the National Academy of Sciences*, 107, 2010, 10827-10832.

[32] Von Neumann, J., "Zur Theorie der Gesellschaftsspiele," *Mathematische Annalen*, 100, 1928, 295-320. English translation by Bargman, S., "On the Theory of Games of Strategy," In Tucker, A., and R.D. Luce (eds.), *Contributions to the Theory of Games*, Volume IV, Princeton University Press, 1955, 13-42.

[33] Von Neumann, J., and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, 1944.